An Analysis of the “Elite” Users on Yelp.com

Kevin Crain Kevin Heh Johhny Winston

December 2014

1 Introduction

Yelp.com is a popular website and mobile app that allows users to both read and post reviews and ratings of restaurants and business. It has a substantial user-base as well, with over 130 million visitors per month. Most customers of the site are casual users, occasionally searching the site to find a good restaurant in their area or expressing their opinions of a place they recently visited. However, there is a very particular set of users of the service which Yelp like to call their “Elite” users. According to Yelp, their Elite users are supposed to be a small group of in-the-know users who have a large impact on their local community. For example, Yelp claims that these users reveal the hot spots for fellow locals, serve as city ambassadors, and have sway in the community. [1]

The purpose of our project is to investigate properties of Yelp’s Elite users. For this paper, we will focus on several of Yelp’s primary claims about their Elite users. First, Yelp states that its Elite users have high connectivity, which means that they are connected with many other users and interact often with members of their Yelp community. Second, Yelp claims that its Elite users make up the “true heart of the Yelp community.” Third, Yelp claims that its users have high contribution, which means that the user has made a large impact on the site with meaningful and high-quality reviews. [1]

The first goal of our project is to analyze whether the above claims about Yelp’s Elite users are valid. For this, we will specify several characteristics which we expect Elite users to have based on these claims. We will then perform analyses on Yelp’s dataset in order to determine whether these properties are truly represented among the Elite users. The secondary goal of our project is to find which properties are most indicative of Elite status on Yelp. The analyses for the first goal can be used for this purpose as well. This kind of information may be useful for those who are interested in becoming Elite members on Yelp. In order to become a member of the “Elite squad,” a user must go through an application process. Despite the suggestions presented above, Yelp doesn’t provide any specific criteria on exactly what characteristics a user must have to become Elite. The mystery behind the selection process for Elite users is well-documented. [2]

2 Prior Work

Our analysis involves several standard network analysis algorithms. One is the PageRank algorithm first proposed by Brin and Page in “The anatomy of a large-scale hypertextual Web search engine” (1998) [3] This algorithm was originally intended to measure the importance of web pages in web search, but we utilize it to measure the importance of nodes in our network. Another algorithm that we use is the Clauset-Newman-Moore community detection algorithm proposed by Clauset et al. in “Finding community structure in very large networks” (2004) [4]. We utilize this algorithm for finding communities within our networks so as to see how well certain nodes are connected to various communities. Our project also utilizes an algorithm for finding betweenness centrality proposed by Brandes in “A Faster Algorithm for Betweenness Centrality” (2001) [5]. We use this algorithm to determine how “central” a node is in our network structures.

3 Methods and Algorithms

The dataset we used comes from the data provided for the Yelp Dataset Challenge. This dataset consists of about 250 thousand users and about 40 thousand businesses from the area around Phoenix, AZ. The dataset includes just over 1.1 million reviews, and there are also connections between friends on Yelp which make up a social graph containing about 950 thousand edges. An edge exists between two nodes if the users represented by those nodes are friends on Yelp.

3.1 Social Network Analysis

The main part of our project involves analyzing Elite users on Yelp’s social network. However, the social graph is undirected and rather large, with 250 thousand nodes and 950 thousand edges. The size of the network means some of our algorithms take an unreasonable amount of time to run. As a result, we decided to run our analyses on a subgraph on the social graph. One approach for this would have been to take a random sample of nodes in the graph. The problem with this approach is that there are a large proportion of nodes which have no edges to other nodes or which are part of very small connected components. As such, taking a random subset of the nodes would destroy much of the structure of the network. Because of this, we instead decided to take the largest connected component of the social graph, and then take a random subset of the nodes in this connected component. We then would keep the edges that exist between these nodes. In the end, our subgraph contained 20 thousand nodes and about 25 thousand edges. This subgraph also has a power-law degree distribution (shown in Figure 1) like the original social graph, and its clustering coefficient of 0.043 is close to that of the original graph, 0.059. As a result, we decided that this subgraph was a suitable replacement for the original and that we would perform all of our analyses on this subgraph.

One of the main characteristics for Elite users specified by Yelp is connectivity. To examine the connectivity of users, we found the degree of each node. We then ordered the users based on their degree and computed the percentage of Elite users found in the top $x\%$ of users for various values of x.

Another claim that we will be examining is that Elite users are the “heart of

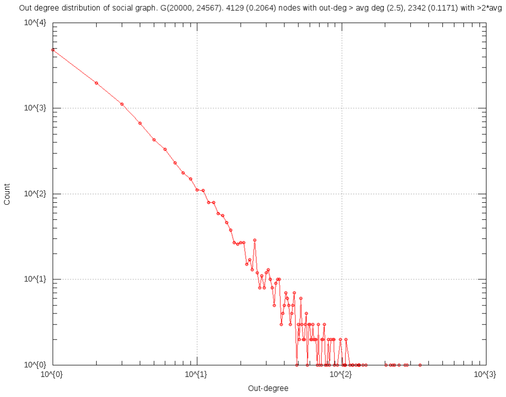

Figure 1: Degree distribution of our social subgraph.

the Yelp community." We will be performing several measurements in order to test this claim. Firstly, we wanted to examine the importance of Elite users to the robustness of the social network. For this, we measured the diameter and the size of the largest connected component of the social graph after removing increasing numbers of nodes from the graph. We compared these values for three different removal strategies: one which involved removing random Elite nodes, one which involved removing random nodes (both Elite and non-Elite), and one which involved removing only non-Elite nodes. Second, we measured the betweenness centrality of each node to determine how central each node is in the network. [5] Third, we found the PageRank scores of every node in the graph to determine the "importance" of each node. [3] Fourth, we found the number of distinct communities that each node is directly connected to. For this, we first ran Clauset-Newman-Moore community detection on the network, keeping only communities with at least 10 nodes. [4] Then for each node, we found the number of distinct communities represented by its neighbors. For each of the latter three properties, we also ordered users based on the property, and computed the percentage of Elite users found in the top $x\%$ of users (in terms of that property) for various values of x.

The last characteristic that we wanted to test was contribution. For this, we decided to find the number of reviews written by each user because writing reviews is the primary method in which users contribute to Yelp, and it is also the primary source of value for the site. For this property, we also ordered the users based on the number of reviews written and computed the percentage of Elite users found in the top $x\%$ of users for various values of x.

If Yelp's claims about their Elite users were valid, then the graph should not be very robust to removing Elite users, meaning that the diameter should increase much more and the size of the largest connected component should shrink much more when Elite nodes are removed than when random or non-Elite nodes are removed. For each of the other properties, the Elite users should be significantly overrepresented in the top $x\%$ of users for the claims to be true.

3.2 Taste Network Creation and Analysis

In addition to using the social graph to analyze Elite users, we thought it would be interesting to look at how users on Yelp may be related by their "tastes." We decided to take a broad definition of "taste," meaning we included both restaurant and non-restaurant business reviews. After all, even if the only business that two users have both rated is a flower shop, this provides us with at least some information about their mutual likes or dislikes. From this graph, we can test whether Elite users make up the heart of the Yelp community in terms of their tastes, rather than their social connections. After creating this network, we analyzed the same properties presented in 3.1 for testing the "heart of the community" claim.

For the taste network, nodes are Yelp users and edges are formed between users who appear to have similar tastes. We began by formulating the dataset as a bipartite graph between users and businesses. In the bipartite graph an edge is formed between a user and a business if the user rated that business. We then "folded" the bipartite graph by removing the business nodes, resulting in a user-user only graph. In doing so, we only wanted to keep edges between users who are similar.

At this point, we faced the challenge of defining a similarity measurement using the given data. For simplicity, ratings from 1-3 were considered "negative," while a 4 or 5 meant "positive." We initially wanted to use an existing similarity measure such as Jaccard similarity ($Jaccard(S_1, S_2) = |S_1 \cap S_2|/|S_1 \cup S_2|$), but these do not capture our knowledge both positive and negative reviews. In the end, we used our own metric of similarity to create our taste graph, which is defined as follows.

Let X be a user. X_{pos} is the set of businesses that X rated positively; analogously, X_{neg} is the set of businesses X rated negatively. Also suppose that a user Y is defined analogously. Let $XY_{same} = \{X_{pos} \cap Y_{pos}\} \cup \{X_{neg} \cap Y_{neg}\}$ and $XY_{dif} = \{X_{pos} \cap Y_{neg}\} \cup \{X_{neg} \cap Y_{pos}\}$. Intuitively, XY_{same} is the set of businesses that both X and Y rated similarly, XY_{dif} is the set of businesses that they rated differently. Finally, we define the following function over pairs of users:

$$Similarity : \text{User} \times \text{User} \mapsto [0,1] \in \mathbb{R}$$

$$Similarity(X,Y) = \begin{cases} 0 & |XY_{same}| = 0 \\ \frac{|XY_{same}|}{|XY_{same}|+|XY_{dif}|} & otherwise \end{cases}$$

Using this metric of similarity, we took the following approach to building the taste graph. With n users, iterate through all $\binom{n}{2}$ possible pairs. For each pair, calculate their similarity score using the bipartite graph, adding an edge if it surpasses a cut-off. Notice that the time complexity for calculating the similarity between any pair of users is dependent on r, the maximum number of reviews by either user. Since $r << n$, calculating the similarity is effectively $O(1)$ time. The time-complexity of this approach is then $O(n^2)$.

The next step was to identify what similarity cutoff to use to form edges in our taste graph. We started by removing all edges with similarity less than 0.5 and called this our "baseline graph." We then measured the number of edges for various cutoffs from 0.5 to 1, increasing by 0.05 each time. We ran this on a random sample of 20 thousand nodes, keeping existing edges between these nodes, because running this algorithm on the entire network would take an unreasonably long time. Figure 2 shows the number of edges in the "baseline graph" that were kept at each cutoff. From the figure, it's clear that not many edges are removed beyond the value 0.55. Hence, we decided to use 0.55 as the cutoff for forming an edge.

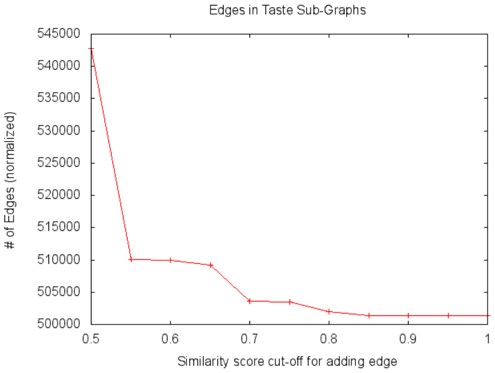

Figure 2: Number of edges in graph for various cutoffs.

4 Results

4.1 Social Network Results

For the social network, we start by looking at the results of our robustness analyses. Figures 3 and 4 show the size of the largest connected component of the graph and the diameter graph, respectively, while increasing the number of nodes removed.

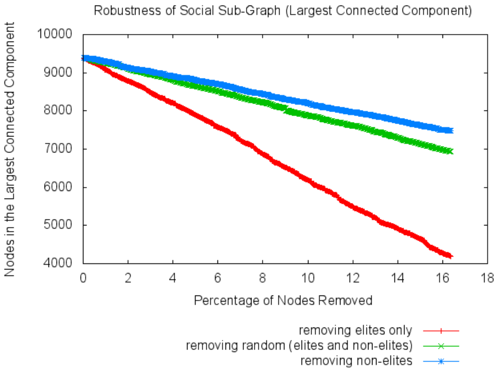

Figure 3: Size of largest CC of social subgraph after removing nodes.

In Figure 3, the size of the largest connected component decreases approximately linearly for all three removal strategies. However, it decreases much more sharply as Elite nodes are removed, compared to when random nodes or random non-Elite nodes are removed. All three plots start at around 9400, but when the percentage of nodes removed reaches 16.35%, which is the percentage of Elite nodes in the graph, the size of the largest CC drops to 4195 when removing only Elite nodes. By comparison, this value drops to 6298 for removing random nodes and 7481 when removing random non-Elite nodes. Hence, the size of the largest CC drops more than twice as quickly when removing Elite as when removing non-Elite nodes.

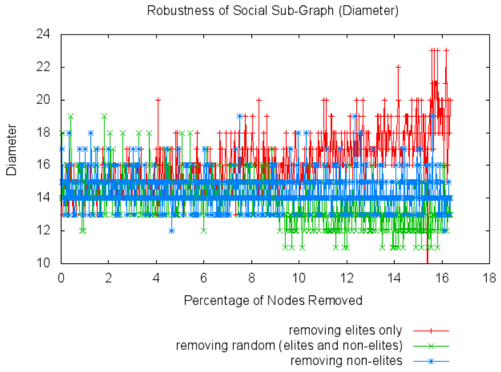

Figure 4: Diameter of social subgraph after removing nodes.

In Figure 4, it's clear that the diameter grows significantly as we continue to remove only Elite nodes. When starting to remove nodes, the diameter ranged from about 13 to 16. But after removing all Elite nodes, the diameter of the graph ranged from about 18 to 23. On the other hand, there is not a significant change in diameter when we remove random nodes or when we remove only non-Elite nodes.

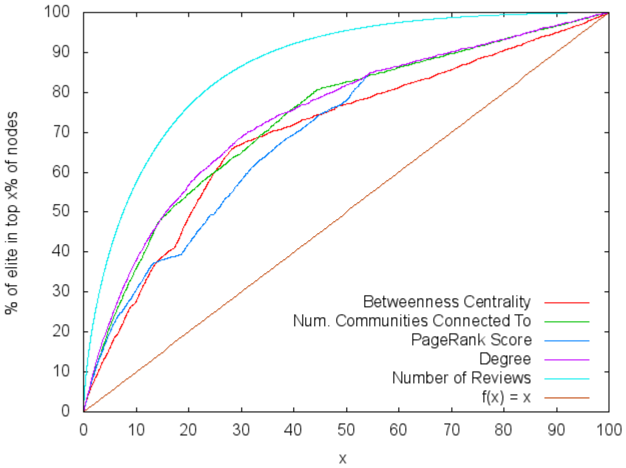

Figure 5: Elite presence within top $x\%$ of users for various properties on the social network.

The remaining measurements for the social graph are shown in Figure 5. Here, for each property, we found the top $x\%$ of users in terms of that property for various values of x. We then computed the percentage of Elite nodes that were part of this top $x\%$ and plotted that value. For instance, a point (x, y) for the degree property means that $y\%$ of Elite nodes are contained within the top $x\%$ of nodes in terms of degree. As stated earlier, we measured the following properties: degree of each node, betweenness centrality of each node, PageRank score of each node, and the number of distinct communities which a node is directly connected to. The figure also includes the plot where we considered the number of reviews that each user wrote. For reference, we also provide the plot $(f(x) = x)$ that would be expected if Elite users had approximately the same values as other users.

From Figure 5, it's clear that Elite users are overrepresented in the top $x\%$ of users for all of the properties that we measured. This is especially prominent for the property of the number of reviews: out of the top 20% of users in terms of reviews written, nearly 80% of those are Elite. Out of the network properties, it seems that the degree of the user and the number of communities they're connected to have slightly greater representation of Elite users. Meanwhile, the properties of PageRank and betweenness centrality appear to have slightly lower representation of Elite.

The idea that Elites tend to have greater values for these properties is further substantiated by Figure 6, which shows the average values for each of the above properties for Elite nodes, all nodes, and non-Elite nodes. For each property, the average value for Elite users is more than double the average over all users. This difference is especially prominent for number of reviews, as the average Elite writes more than seven times as many reviews as the average user.

Property	Elite Avg	All Avg	non-Elite Avg
Degree	8.58	2.46	1.26
Btwn Centr	35253	9284	4208
PageRank	1.13×10^{-4}	5.0×10^{-5}	3.76×10^{-5}
Communities	1.72	0.68	0.48
Reviews	249.64	36.23	17.86

Figure 6: Average values of properties for Elite, all, and non-Elite nodes on the social network.

4.2 Taste Network Results

The robustness analysis of the taste network showed that removing 8.96% of nodes, which is the percentage of Elite users, decreased the size of the largest CC only slightly. When removing all Elites, the largest CC shrinks from 17,306 to 15,276. By comparison, this value drops to 15,600 when removing the same number of random nodes and 15,639 when removing random non-Elite nodes. Hence, the size of the largest CC drops only slightly more when removing Elites as when removing non-Elites.

Each property examined in Figure 7 was plotted using methods explained in section 4.1. It is clear that Elite users are overrepresented in all the properties that we measured. For all properties, more 60% of Elites are in the top 30% of all nodes. Among the network properties, the PageRank of a user seems to have

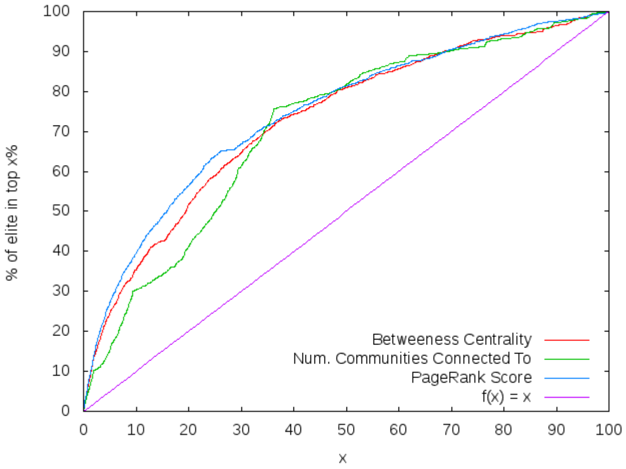

Figure 7: Elite presence within top x% of users for various properties on the taste network.

the greatest representation of Elite users. The number of communities a user is directly connected to and betweenness centrality also have heavy representation of Elites, though less so in the top percentiles. Past the top 35th percentile, all measurements have similar Elite representation. As confirmed in Figure 8, the Elites tend to have significantly greater values for the properties we examined. Moreover, the average Elite had roughly 10 times the betweenness centrality of the average non-Elite.

Property	Elite Avg	All Avg	non-Elite Avg
Btwn Centr	97,797	17,793	9,910
PageRank	1.54×10^{-4}	5.63×10^{-5}	4.68×10^{-5}
Communities	3.10	2.07	1.97

Figure 8: Average values of properties for Elite, all, and non-Elite nodes on the taste network.

5 Conclusions

As stated earlier, the first goal of this project was to test Yelp's claims about characteristics of their Elite users. One claim that they made was that Elite users have high social connectivity. Our results validate this claim with reasonable confidence. On the social network, Elite users are greatly overrepresented among the top degree nodes: 57% of Elite nodes are in the top 20% of nodes in terms of degree. Also, the average degree of Elite nodes (8.58) is significantly higher than the network average (2.46).

Another claim that they made is that Elite users contribute greatly to the site. Our results validate this claim with very high confidence. Elite users are greatly overrepresented among the top reviewers: about 80% of Elite users are in the top 20% of users in terms of reviews written. In addition, the average number of reviews written for Elite users is significantly higher than Yelp users as a whole (250 vs. 36).

The last claim we tested was that the Elite users make up the "heart of the Yelp community." From both a social and taste perspective, our results also validate this claim with reasonable confidence. First of all, the robustness of the social network is much more affected by removing Elite nodes than random nodes (the size of the largest CC decreased more than twice as fast), which suggests that these Elite nodes are important to maintaining the structure of the social network. In addition, the Elite nodes are overrepresented among the top users in terms of being central to the graph (betweenness centrality), being an "important" node in the sense of having high PageRank, and also being directly connected to multiple communities. The fact that Elite nodes have much greater values for these properties than the network as a whole further substantiates the claim.

In the taste network, though, removing Elite nodes had a much smaller effect on the robustness of the graph. This indicates that Elites are more vital to the social network than the taste network. However, Elites are still more important to the structure of the taste network than non-Elites. The importance of Elites was verified by our results for PageRank, betweenness centrality, and direct connectedness to communities. Elites were over-represented among the top percentiles for these three measures. In general, the values of these properties were significantly higher for Elites than non-Elites.

The second goal of our project was to determine which property was the most indicative of Elite status. As mentioned earlier, the Elite users were greatly overrepresented in the top reviewers on Yelp, much more so than for any of the other properties. In adddition, Elite users post on average seven times as many reviews as the average user. This very strong association suggests that, out of the properties we have measured, having a very large numbers of reviews is the best indication of Elite status. This conclusion implies that the number of reviews a user has written would likely be a good predictor of Eliteness. In addition, writing many reviews could be a good objective for those who are looking to become Elite.

6 Future Work

For the taste network, we used our own similarity measure to decide when to connect two nodes. However, it is not a perfect measure of similarity because it does not weight by the number of businesses that two users have both reviewed. For example, if two users have only one restaurant in common and they give a similar review, then our similarity measure would give them a similarity of 1.0. However, if we have two users who have 10 restaurants in common and gave similar reviews for 9 of them, they will only get a similarity of 0.9. Even though the former two users have higher similarity, we actually have much more confidence that the latter two are similar. Hence, our similarity measure may not be ideal in all situations. It would be interesting to explore what other taste networks could be formed using different similarity measurements, and analyze their properties as well.

An extension of our work could also be to use our networks or the properties we measure about the networks to make predictions. For instance, one could use the connections in the social and taste networks, along with the reviews that other users have made, to make predictions about which restaurants a user would enjoy. One could also use these resources to make predictions about which users will become or should become Elite in the future.

7 Acknowledgements

We would like to thank the entire CS 224W staff for providing us with advice and help with the development of this project.

References

[1] "Yelp Elite Squad". http://www.yelp.com/Elite. 2014.

[2] K.V. Brown. "Yelp Elites: Prolific reviewers get perks, VIP treatement". http://www.sfgate.com/restaurants/article/Yelp-Elites-Prolific-reviewers-get-perks-VIP-5664932.php. 2014.

[3] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proc. 7th International World Wide Web Conference, 1998.

[4] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. Physical Review E, 70:066111, 2004.

[5] U. Brandes. A faster algorithm for betweenness centrality. Journal of Mathematical Sociology, 2001.

Contributions:
Kevin Crain: Wrote code/did testing to create taste graph, wrote final report
Kevin Heh: Wrote code for network analysis algorithms, wrote final report
Johnny Winston: Wrote code for user reviews, compiled and plotted data, wrote final report
We all worked to formulate our idea and we helped each other on various aspects.